Final Report

for

Grant No. N00014-90-J-1293

Problems in Survivable Multi-media Networks

C.G. Cassandras, J.F. Kurose, D. Towsley
University of Massachusetts
Amherst, MA 01003

cassandras@ecs.umass.edu, kurose@cs.umass.edu, towsley@cs.umass.edu

1 Introduction

Our research funded under ONR contract N00014-90-J-1293 can be broadly divided into three areas:

1. call admission in high speed networks;
2. optimal routing in the presence of state information;
3. the effects of model uncertainties on routing in networks

These topics will be the subject of the remainder of the technical section. Additional details of our work can be found in the cited technical papers and reports.

2 Call Admission in High Speed Networks

One of the major problems facing designers of high speed networks is the call admission problem. Briefly, the call admission problem is as follows. An application requests the establishment of a session that requires a minimum quality of service (QOS) - typically a limit on the fraction of packets that may be dropped or fraction of packets that may exceed an end-to-end delay bound. Our work has focussed on two aspects of this problem. The first deals with the effects that different algorithms for performing call admission and establishment of a route for the session has on 1) the average delay to set up the call and 2) the fraction of calls that are rejected because of lack of resources. The second deals with the problem regarding the proper choice of a QOS measure for typical applications. We discuss each of these in turr

In future high speed networks (HSNs), significant burdens will be placed on the processing elements in the network since call admission and routing will be more computationally intensive. Thus, the bottleneck will shift from the communication links to the processing elements. The processing delays at these elements are influenced by network parameters such as the routing algorithm, propagation delays, admission control functions (due to QOS requirements), topology, and processing capacities at these elements. We have developed analytic models which capture these parameters and have used them to characterize their influence on the mean call setup time and probability of cal rejection. We considered three sequential routing schemes and two flooding schemes for different network parameters and forms of admission control. The results of our study indicate that call processing capacity associated with call admission affects the probability that a call is rejected significantly but that propagation delays do not. Details are reported in [3, 4].

In addition to the problem Markov decision theory has been successfully used in adaptive routing in traditional circuit-switched networks. Two problems can be identified when extending Markov decision based routing algorithms to future Broadband Integrated Service Digital Networks (B-ISDN's). First, the required computational complexity becomes extremely high in multirate B-ISDN's. Second, a statistical link independence assumption which is made in order to reduce the computational complexity in circuit-switched networks may not be valid in general networks, especially future B-ISDN's, in which network topologies are sparser. We proposed an approach towards adaptive routing in multirate networks using a Markov decision theoretic framework which maintains low computational complexity while still providing quite good routing information. Under this approach, each link is modeled as a birth-death process to reduce the state space size and a policy iteration method from Markov decision theory is iteratively applied to achieve better network performance. Our results show that routing algorithms based on this approach yield better performance than least-load path routing (LLP) without incurring any significant increase in computational complexity. We also investigated the effect of the link independence assumption on the performance of routing schemes. Our results show that routing algorithms yield near optimal performance even when link independence assumptions are severely violated. Details of this work can be found in [4, 5].

Future HSNs are expected to support a wide variety of services such as voice and video and provide a guaranteed quality-of-service (QOS). Traditionally, the computation of user-oriented performance criteria such as the average delay has been carried out via steady state analysis of queueing theoretic models of networks. In this study, we show that steady state analysis is not sufficient for QOS purposes in future high-speed networks as it yields long run performance measures that are not appropriate for applications expected to use HSNs. We propose new QOS criteria based on quality of service over fixed intervals of time and study them via simple queueing models. These new QOS criteria essentially capture the network performance over the short-run and are more accurate indicators of the actual user-perceived quality of applications such as voice or video. In the case of voice, for example, if we are interested in minimizing the occurrence of high loss periods, we find that steady state analysis is optimistic; that it will predict that much higher loads can be supported by a multiplexer than is actually possible. Details of this work can be found in [1, 2]. We consider the problem of routing jobs to parallel queues with identical exponential servers and *unequal finite* buffer capacities. Service rates are state-dependent and non-decreasing with respect to queue lengths. We establish the extremal properties of the *Shortest Non-Full Queue* (SNQ) and the *Longest Non-Full Queue* (LNQ) policies, in systems with concave/convex service rates. Our analysis is based on the weak majorization of joint queue lengths which leads to stochastic orderings of critical performance indices. Moreover, we solve the buffer allocation problem, i.e. the problem of how to distribute a number of buffers among the queues. The two optimal allocation schemes are also 'extreme', in the sense of capacity balancing. Some extensions are also discussed.

3 Optimal Routing

We considered the problem of routing jobs to parallel queues with identical exponential servers and *unequal finite* buffer capacities. Service rates are assumed to be state-dependent and non-decreasing with respect to queue lengths. We established the extremal properties of the *Shortest Non-Full Queue* (SNQ) and the *Longest Non-Full Queue* (LNQ) policies, in systems with concave/convex service rates. Our analysis is based on the weak majorization of joint queue lengths which leads to stochastic orderings of critical performance indices. Moreover, we solved the buffer allocation problem, i.e. the problem of how to distribute a number of buffers among the queues. The two optimal allocation schemes are also 'extreme', in the sense of capacity balancing. Some extensions are also discussed.Details can be found in [8, 12].

We also considered a related problem where the router has no state information. For this case, we established the optimality of the cyclic routing policy for a number of finite buffer systems. Details can be found in [13].

4 Routing in the Face of Uncertainties

We studied the effect of model uncertainties on optimal routing in a system of parallel queues. The uncertainty arises in modelling the service time distribution for the customers (jobs, packets) to be served. For a Poisson arrival process and Bernoulli routing, the optimal mean system delay generally depends on the variance of this distribution. However, as the input traffic load approaches the system capacity, the optimal routing assignment and corresponding mean system delay are shown to converge to a variance-invariant point. An example of a model-independent algorithm using on-line gradient estimation is also included and its performance compared with that of model-based algorithms. Details are found in [11].

In another study, we addressed the problem of routing and admission control of *real-time traffic* in a system where customers must begin service within given deadlines (or complete service within given deadlines), otherwise they are considered *lost*. The performance in such systems is measured by the probability a customer is lost. We developed two approaches for optimal routing and admission control in a system of K parallel servers with a probabilistic routing and admission control scheme. Assuming the availability of a closed-form expression for the probability of loss at each server, the problem is solved under general conditions and properties of the optimal flow allocation are given. However, such closed-form expressions are often unavailable. This motivated the second approach, which involves a gradient-based stochastic optimization algorithm with on-line gradient estimation. The gradient estimation problem for loss probabilities is solved through a recently-developed Smoothed Perturbation Analysis (SPA) technique. The effectiveness of on-line stochastic optimization using this type of gradient estimator is demonstrated by combining the SPA algorithm with a sampling-controlled stochastic optimization algorithm for the aforementioned routing and admission control problem. Details of the two approaches can be found in [14]. Further details on the gradient estimator for this and related problems can be found in [7].

References

[1] Nagarajan, R., Kurose, J., Towsley, D. "Finite-Horizon Statistical Quality of Service Measures for High Speed Networks", submitted to *J. High Speed Networks.*

[2] Nagarajan, R. *Quality-of-Service Issues in High Speed Networks*, Ph.D. thesis, Univ. of Massachusetts, Sept. 1993.

[3] R.-H. Hwang, J.F. Kurose, D. Towsley. " On call processing delay in high speed networks", submitted to *IEEE/ACM Trans. on Networking.*

[4] Hwang, R.-H., *Routing in High-Speed Networks*, Ph.D. thesis, Univ. of Massachusetts, Sept. 1993.

[5] R.-H. Hwang, J.F. Kurose, D. Towsley. "State Dependent Routing for Multirate Loss Networks", *Proc. of GLOBECOM'92*, pp. 565–570, Dec. 1992.

[6] Sparaggis, P., and Cassandras, C.G., "Monotonicity of Cost Functions in a General Class of Queueing Networks", *J. of Discrete Event Dynamic Systems*, Vol. 1, 3, pp. 315-327, 1992.

[7] Wardi, Y., Gong, W-B., Cassandras, C.G., and Kallmes, M.H., "Smoothed Perturbation Analysis for a Class of Piecewise Constant Sample Performance Functions", *J. of Discrete Event Dynamic Systems*, Vol. 1, 4, pp. 393-414, 1992.

[8] Towsley, D., Sparaggis, P., and Cassandras, C.G., "Optimal Routing and Buffer Allocation for a Class of Finite Capacity Queueing Systems", *IEEE Trans. on Automatic Control*, AC-37, 9, pp. 1446-1451, 1992.

[9] Gong, W-B., Yan, A., and Cassandras, C.G., "The M/G/1 Queue with Queue-Length Dependent Arrival Rate", *Stochastic Models*, Vol. 8, 4, pp. 733-741, 1992.

[10] Wardi, Y., Kallmes, M.H., Cassandras, C.G., and Gong, W-B., "Perturbation Analysis Algorithms for Estimating the Derivatives of Occupancy-Related Functions in Serial Queueing Networks", *Annals of Operations Research*, Vol. 39, pp. 269-293, 1992.

[11] Mohanty, B.P., and Cassandras, C.G., "The Effect of Model Uncertainty on Some Optimal Routing Problems", *J. of Optimization Theory and Applic.*, Vol. 77, 2, pp. 257-290, 1993.

[12] Sparaggis, P., Towsley, D., and Cassandras, C.G., "Extremal Properties of the Shortest Non-Full Queue and the Longest Non-Full Queue Policies in Finite Capacity Systems with State-Dependent Service Rates", *J. of Applied Probability*, Vol. 30, pp. 223-236, 1993.

[13] Sparaggis, P., Towsley, D., and Cassandras, C.G., "Routing with Limited State Information in Queueing Systems with Blocking", to appear in *IEEE Trans. on Automatic Control*, 1994.

[14] Kallmes, M.H., and Cassandras, C.G., "Two Approaches to Optimal Routing and Flow Control in Systems with Real-Time Traffic", to appear in *J. of Optimization Theory and Applic.*, 1994.